1

Technical Report 768

Final Report on a National Cross-Validation of the Computerized Adaptive Screening Test (CAST)

Deirdre J. Knapp

Selection and Classification Technical Area

Manpower and Personnel Research Laboratory

U. S. Army

Research Institute for the Behavioral and Social Sciences

November 1987

U. S. ARMY RESEARCH INSTITUTE

FOR THE BEHAVIORAL AND SOCIAL SCIENCES

A Field Operating Agency under the Jurisdiction of the

Deputy Chief of Staff for Personnel

EDGAR M. JOHNSON Technical Director	WM. DARRYL HENDERSON COL, IN Commanding

Technical review by

Jane M. Arabian
Rebecca M. Pliske

Accession For	
NTIS GRA&I	☑
DTIC TAB	□
Unannounced	□
Justification	
By	
Distribution/	
Availability Codes	
Dist	Avail and/or Special
A-1	

DTIC COPY INSPECTED 6

NOTICES

~~DISTRIBUTION: Primary distribution of this report has been made by ARI. Please address correspondence concerning distribution of reports to: U.S. Army Research Institute for the Behavioral and Social Sciences, ATTN: PERI-POT, 5001 Eisenhower Ave., Alexandria, Virginia 22333-5600.~~

FINAL DISPOSITION: This report may be destroyed when it is no longer needed. Please do not return it to the U.S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS
BEFORE COMPLETING FORM

1. REPORT NUMBER
ARI Technical Report 768

2. GOVT ACCESSION NO.
ADA192020

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
FINAL REPORT ON A NATIONAL CROSS-VALIDATION OF THE COMPUTERIZED ADAPTIVE SCREENING TEST (CAST)

5. TYPE OF REPORT & PERIOD COVERED
Final Report
January 1985-March 1986

6. PERFORMING ORG. REPORT NUMBER
--

7. AUTHOR(s)
Deirdre J. Knapp

8. CONTRACT OR GRANT NUMBER(s)
--

9. PERFORMING ORGANIZATION NAME AND ADDRESS
U.S. Army Research Institute for the Behavioral
and Social Sciences
5001 Eisenhower Ave., Alexandria, VA 22333-5600

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
2Q263731A792
2.2.1.H.3

11. CONTROLLING OFFICE NAME AND ADDRESS
U.S. Army Research Institute for the Behavioral
and Social Sciences
5001 Eisenhower Ave., Alexandria, VA 22333-5600

12. REPORT DATE
November 1987

13. NUMBER OF PAGES
31

14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)
--

15. SECURITY CLASS. (of this report)
Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE
--

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

--

18. SUPPLEMENTARY NOTES

--

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)
Recruiting,
Computerized Adaptive Testing (CAT)
~~Computerized~~ Adaptive Screening Test (CAST)
Enlistment Screening Test (EST).

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The Computerized Adaptive Screening Test (CAST) is used by Army recruiters to predict prospective applicants' subsequent performance on the Armed Forces Qualification Test (AFQT). A modified version of the CAST software was used in 60 recruiting stations across the country from January through December 1985 to collect CAST item-level performance information. Screening test data were matched to applicant records from Military Entrance Processing Stations to obtain ASVAB scores and relevant demographic information. The cross-validated,

(Continued)

ARI Technical Report 768

20. ABSTRACT (Continued)

corrected correlation between CAST and AFQT scores is .83. CAST's ability to predict important AFQT performance categories and Army Aptitude Area scores was also examined. Alternative subtest lengths were evaluated and item bank characteristics were described.

Technical Report 768

Final Report on a National Cross-Validation of the Computerized Adaptive Screening Test (CAST)

Deirdre J. Knapp

Selection and Classification Technical Area
Lawrence M. Hanser, Chief

Manpower and Personnel Research Laboratory
Newell K. Eaton, Director

U.S. ARMY RESEARCH INSTITUTE FOR THE BEHAVIORAL AND SOCIAL SCIENCES
5001 Eisenhower Avenue, Alexandria, Virginia 22333-5600

Office, Deputy Chief of Staff for Personnel
Department of the Army

November 1987

Army Project Number
2Q263731A792

Manpower and Personnel

FOREWORD

The Army faces a continuing demand to meet recruiting quality goals. Recent advances in computer technology and psychometric theory have made possible a new type of assessment technique, called computerized adaptive testing (CAT), that can provide accurate ability estimates based on relatively few test items. The Computerized Adaptive Screening Test (CAST) was designed to provide an estimate of a prospect's Armed Forces Qualification Test (AFQT) score at the recruiting station. Recruiters use CAST to help determine whether to send prospects to Military Entrance Processing Stations for further testing and to forecast the various options and benefits for which the prospects will subsequently qualify. This report summarizes analyses from a nation-wide cross-validation of CAST.

This research was conducted under the Manpower and Personnel Research program and contributes to the mission of the Selection and Classification Technical Area to improve the Army's capability to select and classify its applicants using state-of-the-art and fair measures to assess applicant potential. Continuing research and development of CAST is conducted under the sponsorship of the U.S. Army Recruiting Command (USAREC) as outlined in a Memorandum of Understanding dated 29 August 1984 regarding the Army Research Institute/USAREC Research and Development Program. The information in this report was briefed to the Director of Recruiting Operations Directorate, USAREC, on 3 September 1987. The results are being used to further document the acceptability of using CAST as a prescreening tool and to direct future refinement efforts.

EDGAR M. JOHNSON
Technical Director

FINAL REPORT ON A NATIONAL CROSS-VALIDATION OF THE COMPUTERIZED ADAPTIVE SCREENING TEST (CAST)

EXECUTIVE SUMMARY

Requirement:

To evaluate the performance of the Computerized Adaptive Screening Test (CAST) using data from a nationally representative sample of prospective applicants (prospects).

Procedure:

A modified version of the CAST sofware was used in 60 recruiting stations across the country from January through December 1985 so that prospects' CAST performance could be recorded on data diskettes for analysis. CAST performance information was collected from 14,410 examinees. These data were matched to applicant records from the Military Entrance Processing Stations to obtain Armed Services Vocational Aptitude Battery (ASVAB) scores and relevant demographic data for those prospects who went on for further testing. Validity data were examined using regression and cross-tabulation analyses. In addition, the item characteristics of the available Arithmetic Reasoning (AR) and Word Knowledge (WK) item banks were compared to those of the subset of items that were actually administered to the CAST examinees.

Findings:

The correlation between CAST and Armed Forces Qualification Test (AFQT) scores (corrected 1980 Youth Norms) is .79 (N=5,909). When corrected for restriction in range, the correlation is .83. Uncorrected correlations between CAST and Aptitude Area scores range from .64 to .82. For 81% of the examinees, CAST correctly predicted whether or not they would score above the IIIA/IIIB and IIIB/IVA AFQT cutpoints. Most of the WK items available for use were administered more than 15 times during this data collection (63 out of 78). Only 54 of the 225 AR items were administered at least 15 times in this sample of examinees. The item characteristics of the WK item pool are more desirable than the characteristics of the AR item pool; however, both pools meet minimum psychometric standards. Alternative subtest lengths were evaluated using multiple correlation and administration time estimates. There is no compelling evidence for altering the current subtest length at this time.

Utilization of Findings:

This report will be used by the U.S. Army Recruiting Command to justify continued use of CAST as an enlistment screening test.

FINAL REPORT ON A NATIONAL CROSS-VALIDATION OF THE COMPUTERIZED ADAPTIVE SCREENING TEST (CAST)

CONTENTS

CONTENTS (Continued)

Page

LIST OF TABLES

LIST OF FIGURES

FINAL REPORT ON A NATIONAL CROSS-VALIDATION OF THE COMPUTERIZED ADAPTIVE SCREENING TEST (CAST)

INTRODUCTION

The Computerized Adaptive Screening Test (CAST) was designed by the Navy Personnel Research and Development Center (NPRDC) with funding from the Army Research Institute (ARI) to provide a prediction of prospective recruits' Armed Forces Qualification Test (AFQT) scores at recruiting stations. The purpose of this report is to summarize information about CAST that has been obtained through a large scale data collection effort conducted in 1985. The report begins with a brief review of CAST's history and concludes with a review of planned modifications of the test.

History

Applicants for the U.S. armed services are required to take the Armed Services Vocational Aptitude Battery (ASVAB). The paper-and-pencil ASVAB is composed of ten subtests and requires approximately three and one-half hours to administer. The subtest scores are combined to create a variety of composite scores that are used for the selection and classification of enlisted personnel. AFQT scores are currently computed by summing the word knowledge (WK), arithmetic reasoning (AR), paragraph comprehension (PC), and the numerical operations (NO) subtest scores as follows: WK+AR+PC+1/2 NO. The AFQT score is intended to be a measure of an individual's trainability and is used to assess eligibility for enlistment and special benefits. Each service uses unique aptitude area composites to determine eligibility for specific military occupations.

To facilitate the recruiting process, recruiters require information regarding a prospect's probable performance on AFQT. In the late 1970's, the Enlistment Screening Test (EST) was made available to recruiters in all of the armed services (Mathews & Ree, 1982). EST is a traditional paper-and-pencil test that contains 48 items which are similar in content to items in the ASVAB's WK, AR, and PC subtests. In 1984, CAST was made available to Army recruiters. Advantages of CAST over EST include less test administration time and reduced administrative burden for the recruiter. More detailed discussions of why recruiters use screening tests, how they use the tests, and the differences between EST and CAST are presented in earlier CAST reports (Baker, Rafacz, & Sands, 1984; Knapp & Pliske, 1986; Knapp, Pliske, & Elig, 1987; Pliske, Gade, & Johnson, 1984; Sands & Gade, 1983).

Description

CAST is composed of two subtests: WK and AR. It is an adaptive test based on item response theory (Lord, 1980). Thus, the test items administered to a given examinee are selected on the basis of that examinee's estimated ability level (known as theta). There are 78 items in the WK item bank and 225 items in the AR item bank. All CAST items are multiple choice with a maximum of five response alternatives. CAST uses the Bayesian sequential scoring procedure discussed by Jensema (1977) to score and select subsequent items for administration. The item selection procedure incorporates an element of randomization that was intended to reduce item exposure.

CAST is currently administered on the Joint Optical Information Network (JOIN) microcomputer system. JOIN was designed for the U.S. Army Recruiting Command (USAREC) to serve a number of functions at recruiting stations and Military Entrance Processing Stations. The system has 47K of memory available for applications programming.

Development

The item pools for CAST were developed and calibrated by researchers at the University of Minnesota (cf. Moreno, Wetzel, McBride, & Weiss, 1984) for an experimental version of a computerized adaptive ASVAB (CAT ASVAB). The items were drawn from four separate calibration efforts. One-half of the 78 WK items were calibrated on a sample of 677 Marine recruits who took the items via computer. The remaining WK items were calibrated using a sample of approximately 1,300 Marine recruits who took the items using paper and pencil. One hundred and forty-eight of the AR items were calibrated on a sample of Air Force recruits ranging in number from 819 to 1,040 examinees per item. These items were computer-administered. The remaining 77 AR items were calibrated on a sample of 4,100 Navy and Marine recruits using a paper and pencil item administration. All CAST items were calibrated using a three-parameter logistic ogive item response model (Birnbaum, 1968).

Moreno et al. (1984) provided a de facto pilot test of CAST in their research which examined the relationship between corresponding ASVAB and CAT ASVAB subtests. These researchers administered CAT versions of the WK, AR, and PC subtests to 270 male Marine recruits at the Marine Corps Recruit Depot in San Diego, CA. The WK and AR subtest item banks were the same as those described above. The data from this pilot test yielded a correlation of .87 between the three optimally-weighted CAT ASVAB subtests and ASVAB AFQT.

Because the Moreno et al. (1984) data indicated that the PC subtest did not contribute a significant amount of predictive power beyond that provided by the WK and AR subtests, and because the PC subtest items required an inordinate amount of time to administer, this subtest was not incorporated into CAST. Note that an NO subtest was not considered because it is a speeded test that does not lend itself to an adaptive testing format and because it would require precise time limits. Thus, only WK and AR items were administered to the Army applicants who participated in CAST's field test at the Los Angeles Military Entrance Processing Station (Sands & Gade, 1983). Specifically, 20 WK and 15 AR items were administered adaptively to 312 examinees on an APPLE-II microcomputer. Multiple correlation coefficients were computed for each of the 300 possible combinations of subtest lengths. Examination of these results, in light of judgments regarding the probable administration time of the various subtest lengths, led to the recommendation that the operational CAST be terminated following the administration of 10 WK and 5 AR items. The multiple correlation between this optimally-weighted subtest score combination and actual AFQT scores was .85.

Early Cross-Validation Evidence

Army recruiting stations in the midwestern region of the U.S. provided CAST cross-validation data during January and February of 1984 (Pliske et al., 1984). At this point in time, CAST was fully operational in only this region of the country. The CAST scores provided by participating recruiting stations were matched to ASVAB records available from the Military Entrance Processing Command (MEPCOM). CAST and ASVAB data were available for 1,962 individuals. The bivariate correlation between these CAST and AFQT sccres was .80.

Purpose of Present Investigation

This project had several goals. The first goal was to provide a comprehensive evaluation of the prediction equation that had originally been incorporated into CAST. Recall that the multiple regression equation combines the final WK and AR theta estimates to produce a predicted AFQT percentile score. The second goal was to compute a new prediction equation and evaluate its operation. A third goal was to describe the operational nature of CAST in terms of administration time and item pool usage. To date, the descriptions of the two CAST item pools have covered all test items. It has been evident, however, that CAST actually administers only a subset of those items. Thus, a more accurate description of the test would focus on the "operational" subset of items. Finally, this project provides the data required to evaluate CAST's utility for predicting performance on the Army's aptitude area composites. When CAST was introduced, the possibility that it might be useful for predicting eligibility for training assignments was raised, however the relevant data were not available at that time (Sands & Gade, 1983).

Preliminary results that were based on analysis of data collected during the first six months of this project have been documented in two reports (Knapp & Pliske, 1986; Knapp et al., 1987).

METHOD

Subjects

CAST performance information was obtained from 14,410 Army prospects. Correct AFQT percentile scores could be obtained for only 41% (n=5,909) of this sample. The primary reason for failure to obtain AFQT scores for everyone is that many of the CAST examinees never went on to take ASVAB. We have only limited information by which we can evaluate the extent to which this validation sample represents the population of Army prospects. Since CAST is a screening test, the most obvious concern is limited variation in the CAST performance of individuals for whom AFQT scores are available. The mean CAST score (i.e., predicted AFQT percentile score) for the larger unrestricted sample is 39 (SD=20.6) whereas the mean CAST score in the validation sample is 45 (SD=17.9) indicating that such concern is justified. Fortunately, the availability of a good estimate of the population standard deviation (i.e., the SD of the unrestricted sample) permits correction of validity estimates for restriction in range.

Of additional concern regarding the validation sample is the extent to which it represents the population of Army prospects with respect to demographic characteristics. Demographic data are available only for the validation sample so the adequacy of the sample must be inferred on the basis of the sample selection procedure and a priori expectations. The characteristics of the validation sample are summarized in Table 1. This information portrays a reasonable picture of the prospect population. It should be noted that the 60 recruiting stations that participated in the data collection effort were selected to be representative of all Army recruiting stations in terms of geographical location and population density. The sampled stations were also selected to ensure that a relatively large number of black prospects would be included. Indeed, the large percentage of black prospects in the validation sample is the only aspect of the sample which appears at odds with expectations regarding the relevant population.

Table 1

National CAST Cross-Validation (January - December 1985) Sample Description

Sample Size	5,909
Sex	82% Male 18% Female
Race	58% White 38% Black 4% Other

Age	Mean = 20; SD = 3.59
	Median = 19
	Mode = 18

Component	86% Regular Army 14% Army Reserve
AFQT Category (From ASVAB)	24% I and II 17% IIIA 30% IIIB 29% IVA and V

Procedure

Currently, the JOIN system is programmed to record each examinee's name and CAST score onto a "Prospect Data" diskette that the recruiter keeps for his or her own use. A modified version of the CAST software was designed to collect more detailed information onto special data collection diskettes that were sent to ARI for analysis. Information recorded on the diskettes included the identification number of each test item administered to the examinee, the examinee's answer to each item, the time it took for the examinee to read and answer each item, and the examinee's social security number (SSN). The software was also changed so that the prospects would respond to five more items per subtest than are actually used to compute the operational test score.

At the end of each month, during the 12 month data collection period, personnel at each of the 60 participating recruiting stations forwarded the data collection diskettes to ARI. The information on these diskettes was uploaded to a mainframe computer system where it was put into a format that permitted it to be matched to MEPCOM records. MEPCOM records were also provided to ARI on a monthly basis. These records contained not only the subsequent ASVAB (AFQT and other composite) scores but also demographic information for each examinee.

Analyses

The large amount of validation data available from this effort permitted the cross-validation of CAST's original prediction equation and the development and cross-validation of new prediction equations. A new prediction equation was incorporated into the CAST software in 1986. The performance of this revised algorithm is evaluated in terms of its ability to make linear point and category predictions.

In 1986, MEPCOM revised the tables that are used to convert raw AFQT scores to percentile scores. Accordingly, all AFQT scores for the examinees in this investigation were converted to percentile scores using the revised conversion tables. This procedure resulted in the loss of a small number of cases due to insufficient information required to perform the score conversion. The revised AFQT conversion tables also affected the performance of the CAST prediction equation. The impact of this change will be described.

In addition to evaluating CAST's ability to predict AFQT performance, CAST's relationship to Army aptitude area scores will be described. These analyses are intended to provide Army policy-makers with information that would help them evaluate additional uses for CAST.

Finally, the operational nature of the test will be more fully described. Before this data collection effort began, there was very little information regarding administration time and item pool utilization. Analyses reported herein compare the percentage of items available with the percentage of items actually used in operational testing, and describe the psychometric charateristics of these items.

RESULTS

Original Prediction Equation

The correlation between CAST scores derived using the original prediction equation and revised AFQT percentile scores (1980 Youth Norms) is .79. When corrected for restriction in the range of CAST scores, the validity estimate is .83. This estimate is somewhat lower than the estimate provided in Knapp and Pliske (1986) which was .82 (uncorrected). The drop in validity does not appear to be caused by statistical artifacts (e.g., differences in score variance) and it it is too minimal to be of concern.

Developing a New Prediction Equation

To develop a new prediction equation, the data base (n=5,929) was divided into a development sample and a cross-validation sample. Seventy percent (n=4,166) of the examinees were included in the development sample and the remaining examinees (n=1,763) comprised the cross-validation sample. Examinees were selected for each sample on the basis of the last digit of their SSNs.

AFQT scores were regressed on final WK and AR theta values in the development sample. The multiple correlation was .79. The resulting subtest weights were used to compute CAST scores for the cross-validation sample. The bivariate correlation between these CAST scores and AFQT percentile scores was .80. The lack of shrinkage is likely due to the fact that the equation was developed using a sample large enough to provide stable estimates of the regression weights and intercept. This revised regression equation was incorporated into the operational CAST in late 1986.

Once corrected AFQT 1980 Youth norms became available, the procedures described above were used to develop a third regression equation. Although the resulting subtest weights, the multiple correlation, and the standard error of estimate were the same (within rounding error), the intercept was almost 2 points higher. Thus, the prediction equation currently incorporated into CAST yields AFQT score predictions that tend to be a couple of points too low across most of the score range.

Validity of CAST's Point Predictions

Table 2 shows the uncorrected and corrected validity estimates for CAST. These estimates were derived for the entire sample and for selected subgroups of the sample. Figure 1 depicts the regression of AFQT scores onto CAST scores for the total sample. This figure illustrates CAST's tendency to underpredict performance on AFQT. The standard error of estimate associated with this regression is 14 points.

The values in Table 2 show that differences in validity across racial and gender subgroups are slight. Statistical tests for subgroup differences in regression lines confirmed the results reported in Knapp et al. (1987). That is, the AFQT performance of black examinees tends to be overpredicted

Table 2

Bivariate Correlation Between CAST and AFQT Scores by Race and Sex

Group	n	r	r^{c}[a]
All	5,909	.79	.83
White, Non-Hispanic	3,424	.78	.83
Black	2,244	.69	.80
Hispanic	241	.80	.86
Male	4,835	.80	.84
Female	1,074	.77	.82

[a] Correlations corrected for restriction of range in CAST scores.

relative to white examinees, and the AFQT performance of male examinees tends to be overpredicted relative to female examinees. These differences are minimal and parallel those found with other standardized cognitive ability tests (e.g., Dunbar & Novick, 1985; Hanser & Grafton, 1982; Kallingal, 1971).

Validity of CAST's Category Predictions

With currently available data, it is impossible to provide an accurate portrayal of how successful CAST has been with respect to predicting AFQT category classifications. On the basis of an examinee's CAST score, the recruiter predicts the AFQT category to which the examinee is likely to belong. In one of its Army regulations, USAREC has provided recruiters with a table that can be used to convert CAST scores to probability estimates related to subsequent classification into four AFQT categories (see Pliske, et al., 1984 for a discussion of the development of this table). The extent to which recruiters use this conversion table is unknown. Some recruiters may simply interpret CAST scores at face value. For example, if an examinee's CAST score is 49, the recruiter predicts AFQT category IIIB; whereas if the CAST score is 50, the recruiter predicts AFQT category IIIA. Other recruiters, having noted CAST's tendency to underpredict AFQT performance, might conclude that an examinee with a CAST score of 49 is likely to be in AFQT category IIIA. Thus, there are several ways in which a given recruiter may convert CAST point predictions into category predictions.

Figure 2 shows the pattern of these predictions at two AFQT category cutpoints when the assumption is that CAST scores are interpreted at face

DEPICTION OF CAST REGRESSION

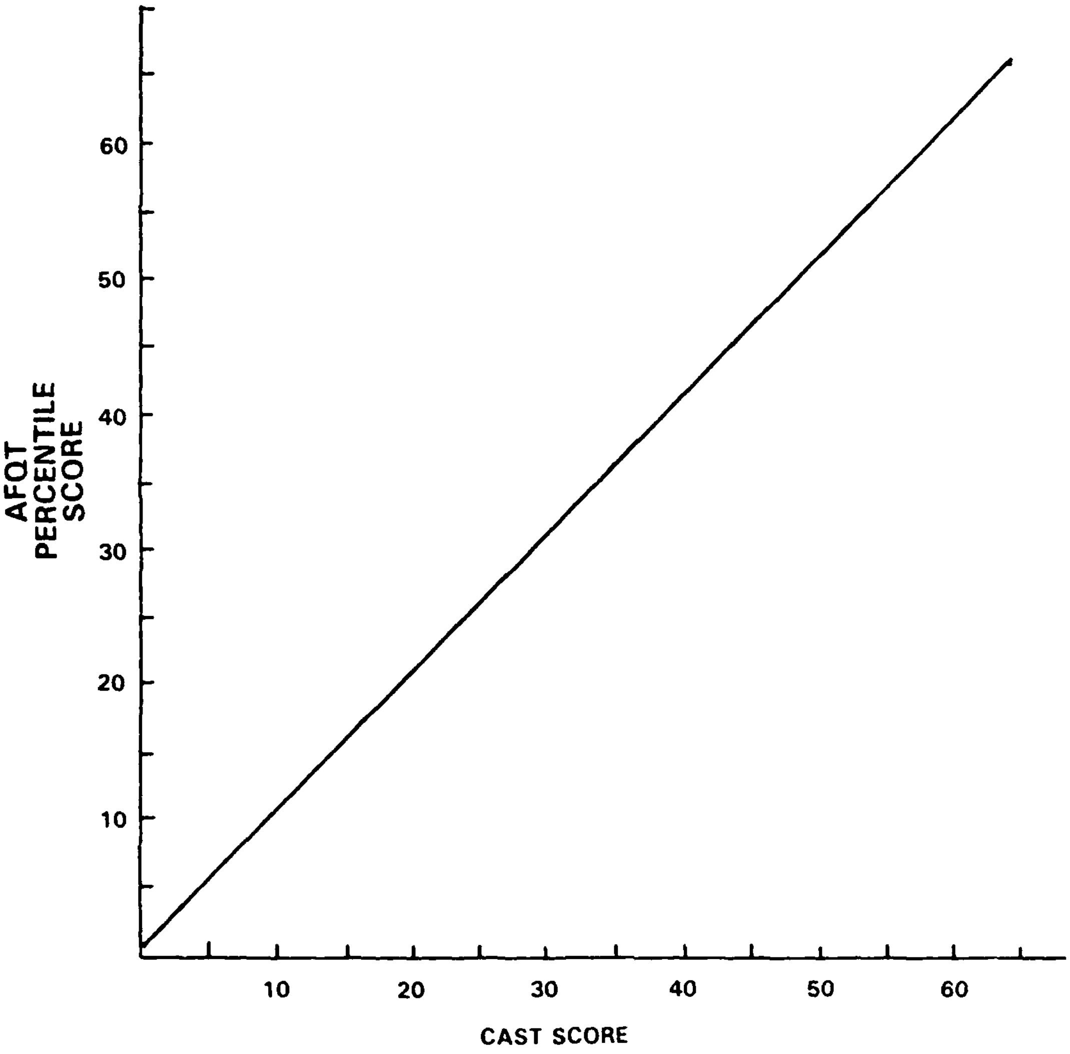

Figure 1

Pattern of CAST Predictions at Two AFQT Category Cutpoints

AFQT Percentile Score		Below 31	31 or Above
	31 or Above	5% Underprediction	66% Hit
	Below 31	15% Hit	14% Overprediction

Figure 2a

CAST SCORE

AFQT Percentile Score		Below 50	50 or Above
	50 or Above	11% Underprediction	30% Hit
	Below 50	51% Hit	8% Overprediction

Figure 2b

CAST SCORE

*Note that the percentages in each table total 100;
AFQT Percentile based on corrected 1980 Norms.

value. A total of 81% of the examinees are correctly classified either above or below both the IIIB/IVA and IIIA/IIIB cutpoints. At the lower end of the ability continuum (Figure 2a), the misclassifications tend to be overpredictions (14%) rather than underpredictions (5%). At the IIIA/IIIP cutpoint, shown in Figure 2b, the opposite is true -- misclassifications tend to be underpredictions (11%) rather than overpredictions (8%). Under the assumption that the conversion table provided for recruiters is used, the overall hit rates remain the same (81%). The only difference is that misclassifications at the IIIB/IVA cutpoint are more likely to be underpredictions (12%) rather than overpredicions (7%).

Relationship to Aptitude Area Scores

Table 3 shows the bivariate correlations between CAST scores and Army aptitude area composite scores. Several of these correlations meet or exceed the size of the correlation between CAST and AFQT. The relationships between CAST and the combat, field artillery, and mechanical maintenance composites, however, are probably too small to be useful.

Table 3

Correlation Between CAST Scores and Army Aptitude Area Composites

Clerical	.82
Combat	.64
Electronic Maintenance	.80
Field Artillery	.65
General Maintenance	.75
Mechanical Maintenance	.65
Operators/Food	.74
Surveillance and Communication	.80
Skilled Technical	.82
General Technical	.81

Note. N = 5,909

Operational Characteristics

Administration Time. Using data from the unrestricted sample to compute time estimates, the mean time required to administer CAST is 16 minutes. This estimate is several minutes higher than that reported in Knapp and Pliske (1986). This is attributable to an error in the reaction time data field that has since been corrected.

Perhaps a more meaningful way to present test administration time information is as follows. Twenty-five percent of the examinees completed CAST in less than 12 minutes, 50% completed CAST in less than 15 minutes and 90% completed CAST within 24 minutes. No steps were taken to trim the time estimates to eliminate random responders or examinees who were interrupted during the course of the test.

Item Banks. Out of the 7o WK items available in the CAST item bank, 63 were administered 15 or more times to the 14,410 CAST examinees. Out of the 225 AR items available for use, only 54 were used 15 or more times in this sample. Thus the "operational" item banks are smaller than the "available" item banks.

Each CAST item has three parameter values associated with it. The first two parameters reflect the discriminability (a-parameter) and the difficulty (b-parameter) of the test item. The third parameter (c-parameter) estimates the probability that the item can be answered correctly by guessing. Urry (1974) outlined the item bank characteristics that would permit efficient and accurate adaptive testing. They are:

1. Item discrimination values as high as possible and no lower than .80.
2. Item difficulty values widely and evenly distributed.
3 Item guessing parameters as low as possible, with .30 as a maximum.
4. There should be a sufficient number of items.

Table 4 shows the distribution of item discrimination values for the available and operational WK item pools. The majority of items in both item pools have discrimination values between 1.0 and 2.0. CAST could probably benefit from a larger number of more discriminating items, however, all items meet the minimum criterion suggested by Urry. Since the majority of available WK items are actually used (81%), it is not surprising that there is little difference in the distribution of discrimination values between the two sets of items.

The distribution of WK item difficulty levels is shown in Table 5. Both the operational and available item pools exhibit a wide range of difficulty values. The available item pool has more easy items (i.e., $\underline{b} < 0$) than difficult items. The distribution of difficulty levels is much more even in the operational item pool but it is still skewed toward very easy items.

Tables 6 and 7 show the distribution of discrimination and difficulty parameters for the available and operational AR item pools. Since the operational item pool contains only 24% of the available items, there are some fairly striking differences between the two sets of items. Although all of the AR items have discrimination values at or above the minimum of .8, 80% of the discrimination values in the available pool of AR items are less than 1.5. In contrast, only 48% of the operational AR items have discrimination values less than 1.5. Thus, there are a large number of AR items that are not used because their discrimination values are relatively low.

Table 4

Distribution of WK Item Discrimination Levels

Available Item Pool

a	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.8 - .9	8	10.3	8	10.3
1.0 - 1.4	45	57.7	53	67.9
1.5 - 1.9	20	25.6	73	93.6
2.0 - 2.4	3	3.8	76	97.4
2.5 - 2.7	2	2.6	78	100.0

Operational Item Pool

a	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.9	3	4.8	3	4.8
1.0 - 1.4	35	55.6	38	60.3
1.5 - 1.9	20	31.7	58	92.1
2.0 - 2.4	3	4.8	61	96.8
2.5 - 2.7	2	3.2	63	100.0

Table 5

Distribution of WK Item Difficulty Levels

Available Item Pool

b	Frequency	Percent	Cumulative Frequency	Cumulative Percent
-2.0 to -1.5	17	21.8	17	21.8
-1.4 to -1.0	11	14.1	28	35.9
-0.9 to -0.5	11	14.1	39	50.0
-0.4 to 0	11	14.1	50	64.1
0.1 to 0.5	9	11.5	59	75.6
0.6 to 1.0	8	10.3	67	85.9
0.9 to 1.5	5	6.4	72	92.3
1.6 to 2.0	6	7.7	78	100.0

Operational Item Pool

b	Frequency	Percent	Cumulative Frequency	Cumulative Percent
-2.0 to -1.5	15	23.8	15	23.8
-1.4 to -1.0	7	11.1	22	34.9
-0.9 to -0.5	6	9.5	28	44.4
-0.4 to 0	8	12.7	36	57.1
0.1 to 0.5	8	12.7	44	69.8
0.6 to 1.0	8	12.7	52	82.5
0.9 to 1.5	5	7.9	57	90.5
1.6 to 2.0	6	9.5	63	100.0

Table 6

Distribution of AR Item Discrimination Levels

Available Item Pool

a	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.7 - .9	56	24.9	56	24.9
1 - 1.4	125	55.6	181	80.4
1.5 - 1.9	37	16.4	218	96.9
2.0	7	3.1	225	100.0

Operational Item Pool

a	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.7 - .9	7	13.0	7	13.0
1 - 1.4	19	35.2	26	48.1
1.5 - 1.9	21	38.9	47	87.0
2.0	7	13.0	54	100.0

Table 7

Distribution of AR Item Difficulty Levels

Available Item Pool

b	Frequency	Percent	Cumulative Frequency	Cumulative Percent
-2.0 to -1.5	0	0	0	0
-1.4 to -1.0	7	3.1	7	3.1
-0.9 to -0.5	23	10.2	30	13.3
-0.4 to 0	22	9.8	52	23.1
0.1 to 0.5	45	20.0	97	43.1
0.6 to 1.0	63	28.0	160	71.1
0.9 to 1.5	39	17.3	199	88.4
1.6 to 2.0	26	11.6	225	100.0

Operational Item Pool

b	Frequency	Percent	Cumulative Frequency	Cumulative Percent
-2.0 to -1.5	0	0	0	0
-1.4 to -1.0	7	13.0	7	13.0
-0.9 to -0.5	12	22.2	19	35.2
-0.4 to 0	5	9.3	24	44.4
0.1 to 0.5	9	16.7	33	61.1
0.6 to 1.0	10	18.5	43	79.6
0.9 to 1.5	7	13.0	50	92.6
1.6 to 2.0	4	7.4	54	100.0

Looking at Table 7 one can see that there is a smaller range of difficulty covered by the AR items than is desirable. In fact, the simplest level of difficulty (b < -1.5) is not represented at all. Only 23% of the items in the available AR item pool have difficulty values less than zero. In the operational item pool, this percentage rises to 44%.

With multiple-choice items, the guessing parameter values are partially determined by the number of response alternatives. Because CAST items generally have five response alternatives, this puts a reasonable upper limit on the size of the c-value. Table 8 shows the distribution of c-parameters for the available WK and AR item pools. The distribution of values for the operational item pools are highly similar so they are not shown. In the WK item pool, approximately 85% of the c-parameter values are below .20 and the majority of the values are between .05 and .15. The AR items tend to have c-parameters that are somewhat higher than the WK items. Most of the values are between .15 and .25. Approximately 58% of the AR c-values are below .20.

Table 8

Distribution of WK and AR Guessing Parameter Values

C	Frequency	Percent	Cumulative Frequency	Cumulative Percent
		WK Item Pool		
.04	2	2.6	2	2.6
.05 - .09	26	33.3	28	35.9
.10 - .14	19	24.4	47	60.3
.15 - .19	19	24.4	66	84.6
.20 - .24	9	11.5	75	96.2
.25 - .26	3	3.8	78	100.0
		AR Item Pool		
.03 - .04	2	0.9	2	0.9
.05 - .09	7	3.1	9	4.0
.10 - .14	23	10.2	32	14.2
.15 - .19	98	43.6	130	57.8
.20 - .24	85	37.8	215	95.6
.25 - .30	10	4.4	225	100.0

Finally, Table 9 shows the correlations between item parameters for the different item banks. In three cases (AR available, WK available, and WK operational), there is a moderate positive correlation between item difficulty and item discrimination. In the AR operational item pool, however, this relationship is quite large (r=.834). Since so few easy items are available, CAST must use items with relatively low discrimination values. Yet there are a large number of difficult items, so CAST uses only those difficult items that are also highly discriminating. This situation has resulted in the high observed correlation between discrimination and difficulty.

Table 9

Correlation Between Item Parameters[a]

	WK Item Pools			AR Item Pools		
	a	b	c	a	b	c
a		.299	.214		.276	.083
b	.236		.170	.834		-.037
c	.224	.254		.118	-.107	

[a] Upper diagonal values are from the available item pools; lower diagonal values are from the operational item pools.

In summary, both the AR and WK item pools meet the minimum standards outlined by Urry (1974). Although the size of the WK item pool is relatively small, the item characteristics are quite acceptable. The item pool could be improved by adding new items that meet or exceed the standards of the old, and that are focused on relatively high difficulty levels. Despite it's size, the AR item pool characteristics are less desirable. The most serious concern is the lack of easy items. The discrimination levels of the items also tend to be lower than desired and the guessing values are a bit high.

Alternative Subtest Lengths

As mentioned earlier, the CAST data collection software recorded theta estimates after each test item was administered. Using this information, we can compute multiple correlation estimates for all possible combinations of subtest lengths up to 15 WK and 10 AR items. Table 10 shows these estimates for combinations of five or more items. As one would expect, larger numbers of test items result in higher validity estimates. One must add several items, however, to produce a noticable increase in validity.

Given that test administration time is also an important consideration in determining test length, Table 11 presents the mean administration times for the subtest length combinations shown in Table 10. The addition of AR items adds appreciably more time to the test than does the addition of WK items.

Table 10

Multiple Correlation Between CAST subtests and AFQT by Subtest Length Combination

		WK										
		5	6	7	8	9	10	11	12	13	14	15
AR	5	76	77	78	79	79	79	80	80	80	80	81
	6	78	78	79	80	80	80	81	81	81	81	81
	7	78	79	80	80	80	81	81	81	82	82	82
	8	79	80	80	81	81	81	82	82	82	82	82
	9	80	80	81	81	81	82	82	82	82	83	83
	10	80	81	81	82	82	82	82	83	83	83	83

Note. Decimal points have been omitted.

Table 11

Mean Test Administration Time (In Minutes) by Subtest Length Combination

		WK										
		5	6	7	8	9	10	11	12	13	14	15
AR	5	14	15	15	15	16	16	17	17	18	18	19
	6	16	16	16	17	17	18	18	19	19	20	20
	7	17	17	18	18	19	19	20	20	20	21	21
	8	18	19	19	20	20	21	21	21	22	22	23
	9	20	20	21	21	21	22	22	23	23	24	24
	10	21	22	22	22	23	23	24	24	25	25	25

Comparison of Tables 10 and 11 shows that a change in the subtest length of CAST is not strongly supported. It would take an average of 25 minutes to administer 10 AR and 15 WK items (the longest subtest length that can be considered from these data). The validity estimate would increase from the current .79 to .83. The standard error of estimate would decrease from 14 points to 13 points. Thus, even the maximum subtest length evaluated here does not yield a particularly substantial increase in validity.

SUMMARY

In January 1985, 60 Army recruiting stations were asked to begin forwarding CAST data to ARI. This data collection effort continued through the end of December 1985. In this paper, the details of the data collection procedures and statistical analyses of the data have been reported. Whereas earlier reports related to this data collection effort focused on data gathered during the first six months of 1985, the analyses reported herein are based on the entire CAST dataset.

The data collection effort in 1985 resulted in a large amount of useful information. It has provided solid evidence of CAST's ability to predict AFQT performance. This evidence was needed to provide a clear justification for using CAST as a tool for screening potential Army applicants. Despite the relatively strong relationship between CAST and AFQT scores, however, this screening test could be refined to better suit the Army's needs. In fact, refinement efforts are currently underway, and these efforts rely heavily on information from the 1985 data base.

Possible immediate changes to CAST include the way in which the results are displayed, the algorithm used to compute AFQT percentile scores, and the length of the two subtests. Each of these areas of potential change will be briefly reviewed.

ARI has suggested several alternative ways to present CAST results to recruiters (Knapp, 1987). Basically, two approaches were considered. One approach is based on the prediction intervals associated with the CAST estimates of AFQT percentile scores. The second approach is based on the estimated probability that an individual with a given CAST score will fall into one of three or four AFQT performance categories. The information needed to program these alternative output displays into the CAST software could only be derived from a data base such as that created in 1985.

These data also allow the computation of a stable and precise algorithm for deriving predicted AFQT percentile scores from the CAST subtest scores. Despite potential changes in the way in which AFQT scores are computed (i.e., replacing the Numerical Operations subtest with Math Knowledge) and previous changes in the derivation of AFQT percentile scores, this data base can provide the appropriate prediction algorithm as needed. A software change to correct the intercept of the current algorithm and the display of CAST results is pending.

Finally, analyses reported herein related to the changes in predictive ability and test administration time as a function of subtest length have been used to reevaluate the current subtest length of CAST. The addition or deletion of several WK items has little effect on either the validity estimate or the testing time. Presently, so few AR items are administered that it would be very risky to consider reducing their number. Adding AR items significantly increases testing time (1-2 minutes per item) with little payoff in terms of increased predictive accuracy. Thus these data seem to justify the current CAST subtest length. Note, however, that future changes to CAST that affect the internal testing strategy may influence the subtest length issue.

As a result of a major refinement effort that began in 1987, a new version of CAST, to be known as CAST II, will be developed. The focus of this refinement project will be to reconsider the CAST subtest scoring and item selection algorithms and to improve the item banks. As part of this refinement project, another large scale data collection will be required. Item calibration data will be collected from new recruits at Army Reception Battalions and from CAST examinees at recruiting stations. CAST II will be available for operational use in 1988.

The 1985 CAST data base continues to provide information that directs this major refinement effort. The most obvious example is related to the item selection rule and improvement of the item pools. ARI's existing data base confirms that some items are over-used and clearly shows the pattern of item usage. This information will help to determine a more appropriate item selection algorithm and to decide if some test items should be deleted from the item banks.

Assuming that the developmental work for CAST II will result in an enduring internal testing framework, the remaining problem will be to ensure that the item banks and AFQT prediction algorithm are periodically monitored and updated. A special version of the CAST II software will have the capability of collecting data that can be used to accomplish this maintenance function in a relatively unobtrusive manner.

Although adaptive testing is very efficient when compared to traditional testing, it can be quite costly in research and development resources. The primary problem is that each potential test item needs to be administered to close to 2,000 people to provide an adequate assessment of its psychometric properties. Rather than collecting such data all at once, it is possible to collect the data a little at a time. That is, one can embed several non-scored test items into the operational version of the test and record the item response information for future research use.

Thus, the long-term maintenance program calls for the periodic addition of experimental items to the operational CAST II software. These items will be administered in a manner that will be transparent to both the examinees and the recruiters. Data from these items, CAST performance scores, and examinee SSN will be electronically transmitted from recruiting stations to a central data base. As time passes, sufficient data will become available to calibrate the experimental test items. Periodic statistical analysis of these data and examination of operational item usage information will allow regular updating of the item banks. Also at regular intervals, CAST performance scores will be matched to applicant records to verify the relationship between those scores and subsequent AFQT performance.

Not only has the 1985 CAST data collection provided convincing evidence that CAST is a useful screening tool, it also continues to be a rich source of information for CAST refinement efforts. These modifications are intended to result in a test with outstanding psychometric qualities and minimum maintenance requirements. Another important goal is to achieve maximum flexibility to ensure the test's continued usefulness in an ever-changing recruiting environment.

REFERENCES

Baker, H. G., Rafacz, B. A., & Sands, W. A. (1984). Computerized Adaptive Screening Test (CAST): Development for use in military recruiting stations (NPRDC Report No. 84-17). San Diego, CA: Navy Personnel Research and Development Center.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds) Statistical theories of mental test scores. Reading, Mass: Addison-Wesley.

Dunbar, S. B. & Novick, M. R. (1985). On predicting success in training for males and females: Marine Corps clerical specialties and ASVAB forms 6 and 7 (ONR Report No. 85-2). Washington, DC: Office of Naval Research.

Hanser, L. M. & Grafton, F. C. (1982). Predicting job proficiency in the Army: Race, sex, and education (Selection and Classification WP No. 82-1). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Jensema, C. G. (1977). Bayesian tailored testing and the influence of item bank characteristics. Applied Psychological Measurement, 1, 111-120.

Kallingal, A. (1971). The prediction of grades for Black and White students at Michigan State University. Journal of Educational Measurement, 8, 263-266.

Knapp, D. J. (1987). Display of results: Alternatives for the Computerized Adaptive Screening Test (CAST). (Selection and Classification Technical Area Working Paper 87-05). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Knapp, D. J. & Pliske, R. M. (1986). Preliminary report on a national cross-validation of the Computerized Adaptive Screening Test (CAST). (ARI Research Report No. 1430). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A175 767)

Knapp, D. J., Pliske, R. M., & Elig, T. W. (1987). Cross-validation of the Computerized Adaptive Screening Test (CAST): An examination of test fairness. (ARI Technical Report No. 747). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. In press.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Mathews, J. J. & Ree, M. J. (1982). Enlistment Screening Test Forms 81A and 81B; Development and calibration (AFHRL Report No. 81-54). Brooks Air Force Base, Texas: Air Force Human Resources Laboratory.

Moreno, K. E., Wetzel, C. D., McBride, J. R., & Weiss, D. J. (1984). Relationship between corresponding Armed Services Vocational Aptitude Battery (ASVAB) and computerized adaptive test (CAT) subtests. Applied Psychological Measurement, 8, 155-163.

Pliske, R. M., Gade, P. A., & Johnson, R. M. (1984). Cross-Validation of the Computerized Adaptive Screening Test (CAST) (ARI Research Report No. 1372). Alexandria, VA; U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A163 148)

Sands, W. A. & Gade, P. A. (1983). An application of computerized adaptive testing in U.S. Army Recruiting. Journal of Computer-Based Instruction, 10, 87-89.

Urry, V. W. (1974). Approximations to item parameters of mental test models and their uses. Educational and Psychological Measurement, 34, 253-269.